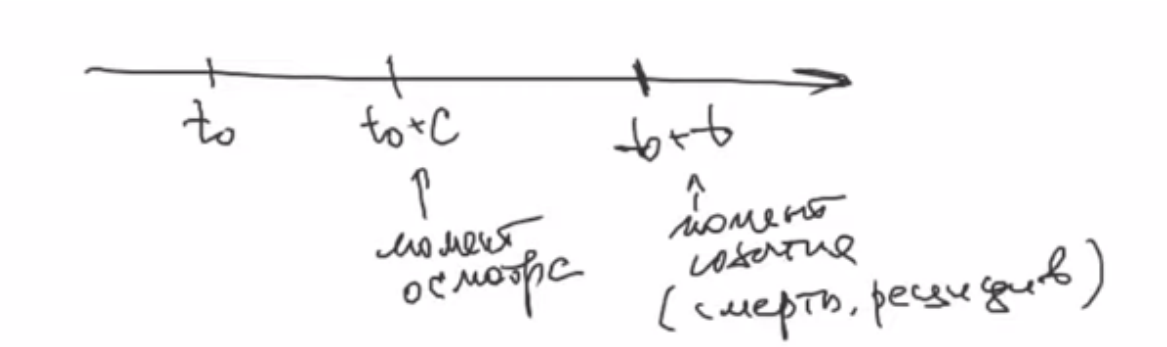
Анализ выживаемости. Продолжение

Вспоминаем предыдущее занятие

Есть пациенты, больные раком. Некоторые умирают, некоторые излечиваются. А некоторые пропадают с радаров: они перестают ходить на ежегодные осмотры к врачу



Для анализа вводим две функции:

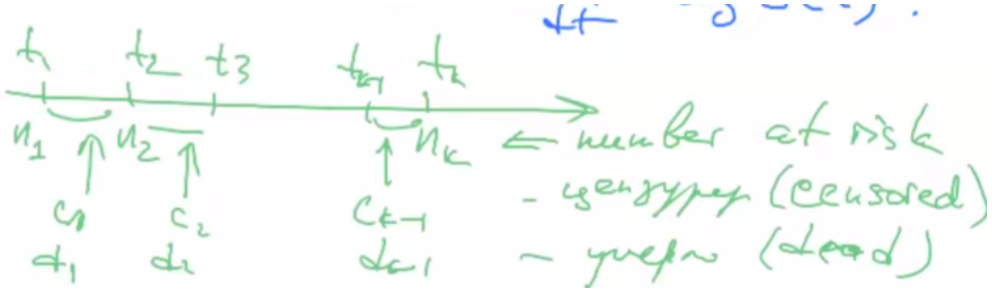
- Survival function: $s(t) = \mathbb{P}\{T \geqslant t\} = 1 - F_T(t)$
- Hazard function: $h(t) = \lim \frac{\mathbb{P}\{t<T\leqslant t+\Delta t \mid T \geqslant t\}}{\Delta t}$. Является некоторым аналогом мгновенной функции.

Свойства:

- $h(t) = -\frac{d}{dt}\log s(t)$

Как оценивать survival function?

Live-table estimate



- n_i - сколько людей есть в i-ом интервале времени
- d_i - сколько людей умерло в i-ом интервале
- c_i - сколько людей наблюдались последний раз в i-ом интервале

Actual assumption:

$$s^*(t) = \prod_{j=1}^{k}\left(1 - \frac{d_j}{n'_j}\right), \quad n'_j = n_j - \frac{c_j}{2}$$

На самом деле непонятно, почему не считать n'_j как $n_j - \frac{c_j+d_j}{2}$ - видимо чисто исторические причины

Kaplan-Meier estimate

Этот подход, в отличие от предыдущего, реально используется на практике.

Пусть t_j - момент смерти j-го пациента. Тогда в качестве оценки s можно взять

$$\hat{s}(t) = \prod_{j=1}^{k} \left(1 - \frac{d_j}{n_j}\right)$$

По сути все отличие от предыдущей формулы в том, что мы вместо n'_j написали n_j. Объясним, почему так можно сделать:

$$\mathbb{P}\left\{T > t_{(k)}\right\} = \mathbb{P}\left\{T > t_{(k)} \mid T > t_{(k)} - \delta\right\} \cdot \mathbb{P}\left\{T > t_{(k)} - \delta \mid T > t_{(k-1)}\right\} \cdot \mathbb{P}\left\{T > t_{(k-1)}\right\}$$

Между красными моментами никто не умирал. Тогда

$$\underbrace{\mathbb{P}\left\{T > t_{(k)} \mid T > t_{(k)} - \delta\right\}}_{1-\mathbb{P}\left\{T=t_{(k)}\right\}=1-\frac{d_k}{n_k}} \cdot \underbrace{\mathbb{P}\left\{T > t_{(k)} - \delta \mid T > t_{(k-1)}\right\}}_{=1} \cdot \mathbb{P}\left\{T > t_{(k-1)}\right\}$$

Расписывая по индукции последнее слагаемое получаем требуемую формулу.

Если посмотреть на формулу, то может возникнуть вопрос: а куда делись цензурируемые события? На самом деле, если положить $c_k = 0$, то получим:

$$(\ldots) = \frac{n_k - d_k}{n_k} \cdot \frac{n_{k-1} - d_{k-1}}{n_{k-1}} \cdot \ldots \cdot \frac{n_1 - d_1}{n_1} = \frac{n_{k-1}}{n_k} \cdot \ldots \cdot \frac{n_2}{n_1} = \frac{n_{k+1}}{n_1}$$

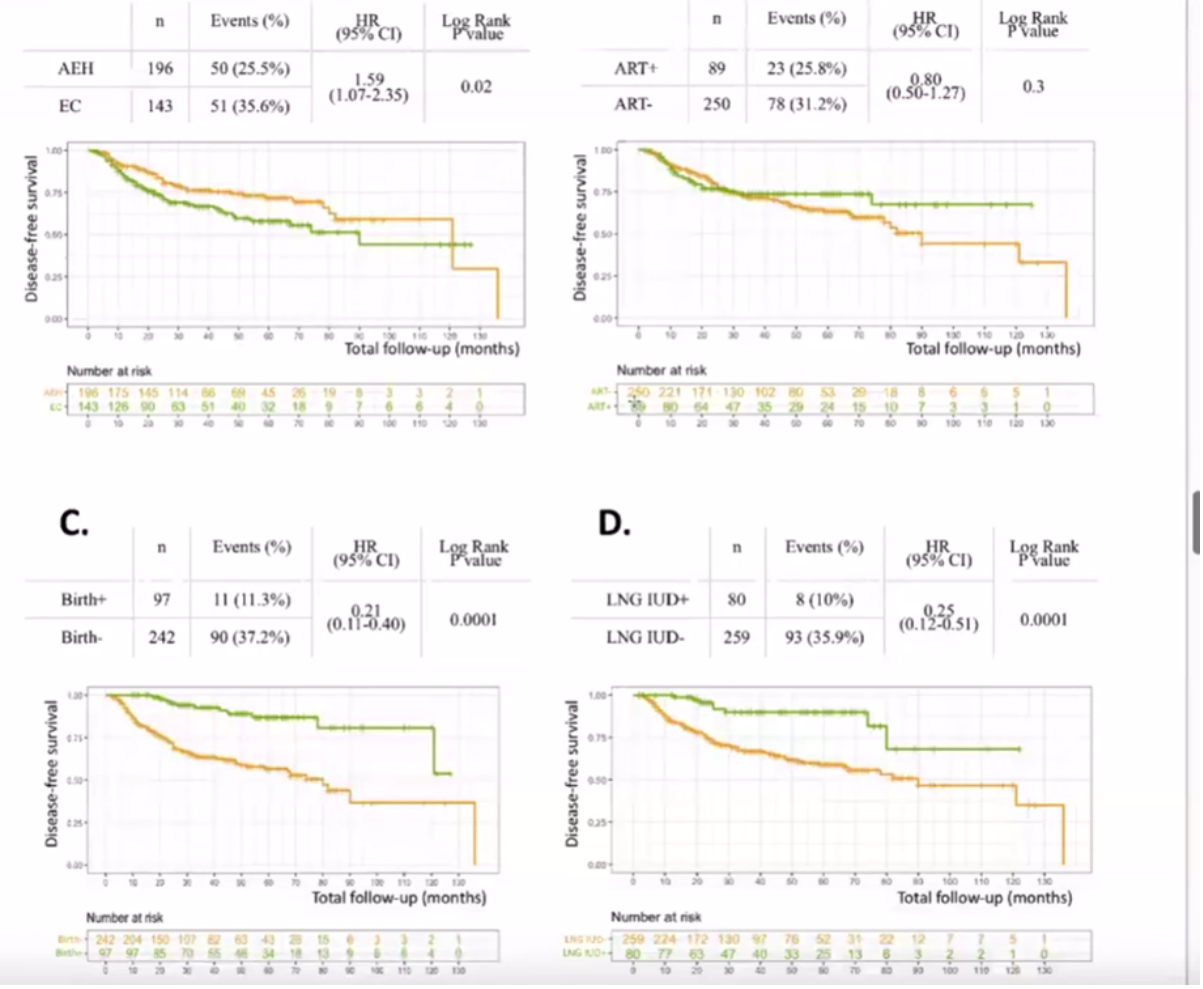
Пример из статьи

Изучается женская болезнь: рак матки. После того, как пациентке сделали операцию, он зачастую хочет родить ребенка. Исследуется вероятность возникновения рецидива (повторного возникновения болезни) в зависимости от некоторых факторов:

- тип диагноза [A]
- роды ребенка проводятся по специальной технологии [B]
- факт рождения ребенка [C]
- применяется специальный курс лечения [D]

Соответственно, мы хотим проверить 4 гипотезы.

Посмотрим на графики. Черточки - это моменты цензуирования. Те моменты, когда кривая прыгает - это либо момент наступления рецидива, либо же момент, когда человек перестает обследоваться.



Оценка hazard function

Еще раз определение:

$$h(t) = \lim \frac{\mathbb{P}\left\{t < T \leqslant t + \Delta t \mid T \geqslant t\right\}}{\Delta t}$$

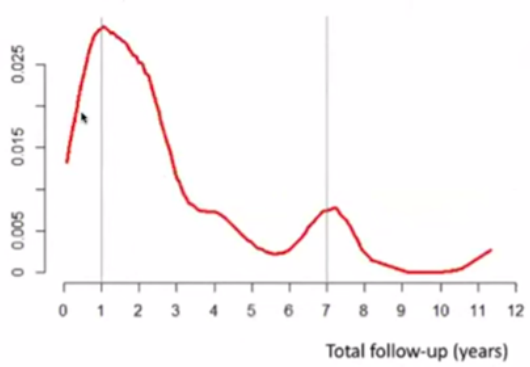
Live-table estimate

$$h^*(t) = \frac{d_j}{n'_j \cdot (t_{j+1} - t_j)}, \ \forall t \in [t_j, t_{j+1}], \quad n'_j = n_j - \frac{d_j + c_j}{2}$$

Kaplan-Meier

$$\hat{h}(t) = \frac{d_j}{n_j \cdot (t_{j+1} - t_j)}$$

Что получилось в статье



Смысл hazard function - вероятность наступления рецидива, если раньше его не случалось. Видно, что для нашей задачи рецидив часто наступает через 1 год и через 7 лет.

Сравнение групп

Есть две группы: 1 и 2. Их hazard functions: h_1, h_2.

Proportional hazard model:

$$h_2(t) = C \cdot h_1(t), \quad C > 0, \neq 1$$

В терминах функции выживаемости: соответствующие функции выживаемости s_1, s_2 не пересекаются во всех точках, где они одновременно не равны нулю или единице

$$h(t) = -\frac{d}{dt} \log s(t) \;\Rightarrow\; s(t) = \exp\left\{-\int_0^t h(u)\,du\right\}$$

$$\Rightarrow\; s_2(t) = \exp\left\{-\int_0^t h_2(u)\,du\right\} = \exp\left\{-\int_0^t C \cdot h_1(t)\,dt\right\} = (s_1(t))^2$$

Если $s_1(t) = (s_1(t))^C \;\Rightarrow\; s_1(t)$ либо 0, либо 1 (а это крайние значения). Значит, при предположении $h_2(t) = C \cdot h_1(t)$ графики $s(t)$ не пересекаются.

Для пациента номер i:

$$h_i(t) = e^{\beta x_i} h_0(t), \quad x_1 = \begin{cases} 0, & \text{в первой группе} \\ 1, & \text{во второй группе} \end{cases}, \quad e^\beta = C, \quad h_0 = h_1$$

По сути получили универсальный вид функции. Функцию h_0 часто называют baseline hazard.

Функцию h_i можно обобщить. Пусть $\overrightarrow{x_i}$ - вектор характеристик i-го пациента. Тогда

$$h_i(t) = e^{\langle \bar{\beta}, \bar{x}_i \rangle} h_0(t)$$

Обычно x_i выбирают так, что когда они равны нулю, мы получаем элемент из первой группы.

Если присмотреться, то мы получили нечто вроде задачи регрессии.

Как оценить β?

Введем аналог функции правдоподобии. Пусть $t_{(1)} < \cdots < t_{(k)}$ - моменты когда кто-то умер (возник рецидив в нашей задаче). Рассмотрим функцию L:

$$L(\bar\beta) = \prod_{j=1}^{n} \mathbb{P}\left\{\text{пациент с характеристикой } \bar x_j \text{ умер в момент времени } t_{(j)} \mid \text{в момент } t_{(j)} \text{ есть} \geqslant 1 \text{ смерть}\right\}$$

$$= \prod_{j=1}^{n} \frac{\mathbb{P}\text{пациент с хар. } \bar x_j \text{ умер в момент } t_{(j)}}{\sum_{s\in R(t_{(j)})} \mathbb{P}\left\{x_s \text{ умер в момент } t_{(j)}\right\}}$$

где $R(t_{(j)}) -$ пациент в зоне риска в момент $t_{(j)} - \delta$

$$(\dots) = \prod_{j=1}^{n} \frac{\mathbb{P}\text{пациент с хар. } \bar x_j \text{ умер в момент } t_{(j)}/\delta}{\sum_{s\in R(t_{(j)})} \mathbb{P}\left\{x_s \text{ умер в } [t_{(j)}-\delta, t_{(j)}]\right\}/\delta}$$

где $R(t_{(j)})$ - пациент в зоне риска в момент $t_{(j)} - \delta$. Далее,

$$L(\bar\beta) = \prod_{j=1}^{n} \frac{\overbrace{\mathbb{P}\text{пациент с хар. } \bar x_j \text{ умер в момент } t_{(j)}}^{\to h_j(t_{(j)})}/\delta}{\sum_{s\in R(t_{(j)})} \underbrace{\mathbb{P}\left\{x_s \text{ умер в } [t_{(j)}-\delta, t_{(j)}]\right\}}_{\to h_s(t_{(j)})}/\delta}$$

$$\to \prod_{j=1}^{n} \frac{h_j(t_{(j)})}{\sum_{s\in R(t_{(j)})} hs(t_{(j)})} = \left\{h_i(t) = e^{\langle\bar\beta, \bar x_i\rangle} h_0(t)\right\}$$

$$= \prod_{j=1}^{n} \frac{e^{\langle\beta, x_j\rangle}}{\sum_{s\in R(t_{(j)})} e^{\langle\beta, x_s\rangle}}$$

$$\to \max_{\beta}$$

Чтобы оценить β надо максимизировать дробь по β. Допустим, мы научились максимизировать. Что делать дальше?

Логранговый критерий (logrank test, score test)

Вещь, которую мы получим ниже, является в некотором смысле фундаментальной. Можно прийти к нему разными способами

Способ 1

$$L(\beta) = \prod_{j=1}^{k} \frac{e^{\beta x_j}}{\sum_{s=1}^{n_j} e^{\beta x_j}}$$

где n_j - количество людей в риске в момент j. Логорифмируем:

$$\log L(\beta) = \beta \sum_{j=1}^{n} x_j - \sum_{j=1}^{n} \log\left(\sum_{s=1}^{n_j} e^{\beta x_j}\right)$$

Обозначим $d_2 := \sum_{j=1}^{n} x_j$ - общее количество умерших людей из второй группы. Вторую сумму распишем как $\sum_{j=1}^{n_j} e^{\beta x_s} = n_{1j} + n_{2j} e^{\beta}$. Поясняющая таблица:

group	**death at j**	**number at rish at j**
1	d_{1j}	n_{1j}
2	d_{2j}	n_{2j}
	$d_i = \sum_j d_{ij}$	

В новых обозначениях получаем:

$$\log L(\beta) = \beta d_2 + \sum_{j=1}^{n} \log(n_{1j} + n_{2j} e^{\beta})$$

Score test проверяют гипотезу

$$\mathcal{H}_0 : \quad \beta = 0$$

Статистика:

$$\frac{\frac{\partial}{\partial \beta} \log L(\beta)\Big|_{\beta=0}}{-\frac{\partial^2}{\partial \beta^2} \log L(\beta)\Big|_{\beta=0}} \sim \mathcal{X}_1^2$$

Если $\beta = 0$, то получается группы одинаковые.

Способ 2

Вывод через гипергеометрическое распределение. Что это такое? Пусть есть N объектов, D из них отмечены. Мы берем n из них. Тогда вероятность, что d из них отмечены:

$$\mathbb{P}\left\{\text{отмеч.} = d\right\} = \frac{C_0^d C_{N-0}^{n-d}}{C_N^n}$$

$$\mathbb{E}[\text{отмеч}] = \frac{n}{N} D$$

Лирическое отступление про пользу гипергеометрического распределения

Допустим у нас есть пруд и мы хотим оценить количество рыбы в нем. Первым шагом мы ловим D рыб и отмечаем их, например вешая на них ленточку. Далее мы ловим n рыб и смотрим сколько среди них отмеченных - получаем d_1. Повторяем эксперимент несколько раз - получаем числа $d_2, \dots d_m$. Теперь мы имеем выборку $d_1, \dots d_m$ и можем максимизировать функцию правдоподобия:

$$\prod_{i=1}^{m} p(d \mid n, D) \to \max_N$$

В нашем случае

- $N = \bar{n}_j = n_{1j} + n_{2j}$
- $D = \bar{d}_j = d_{1j} + d_{2j}$
- $n = n_{1j}$

Тогда

$$\mathbb{P}\left\{\text{количество смертей в первой группе} = d_{1j}\right\} = \frac{C_{\bar{d}_j}^{d_{1j}} C_{\bar{n}_j - \bar{d}_j}^{n_{1j} - d_{1j}}}{C_{\bar{n}_j}^{n_{1j}}}$$

$$\mathbb{E}[\text{количество смертей в первой группе}] = \frac{n_{1j}}{\bar{n}_j} \bar{d}_j$$

Как померить отклонения наблюдаемых отклонений от ожидаемых?

$$U_L = \sum_{j=1}^{n} \left(d_{1j} - \frac{n_{1j}}{\bar{n}_j} \bar{d}_j \right) = d_1 - \sum_{j=1}^{n} \frac{n_{1j}}{\bar{n}_j} \bar{d}_j$$

Можем вычислить некоторые статистики:

$$\mathbb{E}U_L = 0, \quad \operatorname{Var} U_L = \sum_{j=1}^{n} \frac{n_{1j}\, n_{2j}\, \bar{d}_j (\bar{n}_j - \bar{d}_j)}{n_j^2 (n_j - 1)}$$

Утверждение

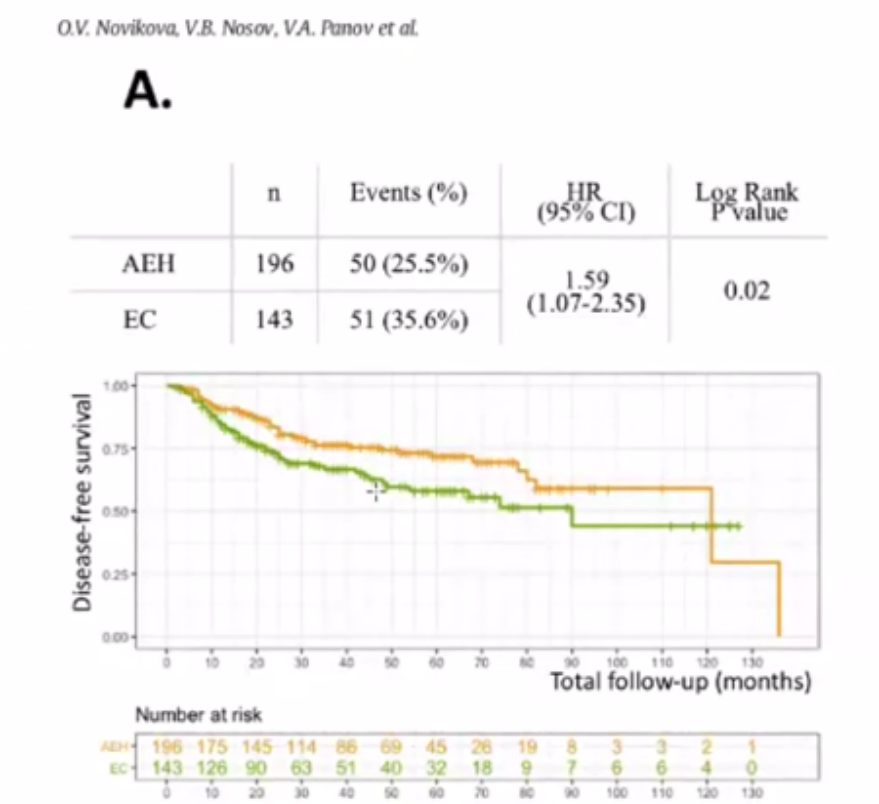
$$\frac{U_L}{\sqrt{\operatorname{Var}} U_L} \to \mathcal{N}(0, 1)$$

Отсюда получаем

$$\frac{U_L^2}{\operatorname{Var} U_k} \to \mathcal{X}_1^2$$

Что отсюда можем получить? Если в каждый момент у нас ровно одна смерть (то есть $\bar{d}_j = 1$), то формула, получаемая из гипергеометрического распределения, полностью совпадает с формулой выше.

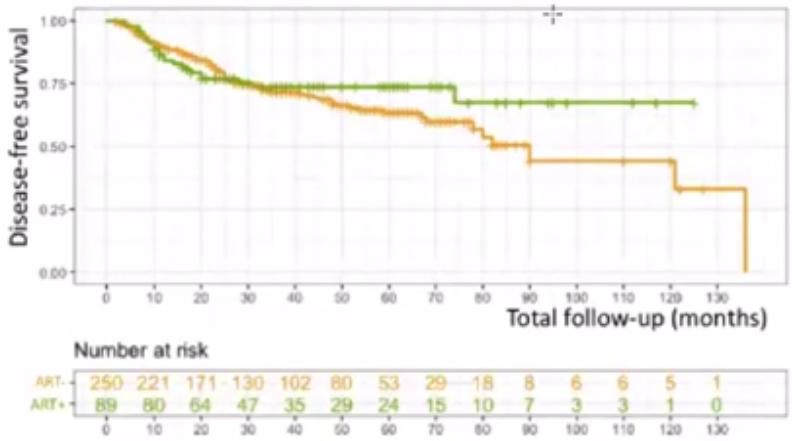
Пример из статьи



Наши функции не пересекаются до 120-го месяца. Вообще это плохо, но вроде как мы можем просто выкинуть эти наблюдения. Уровень значимости p-value=0.02 , что говорит о различии групп. Эти графики были для уровня диагноза и выглядят логично: чем диагноз хуже, тем вероятнее рецидив.

B.

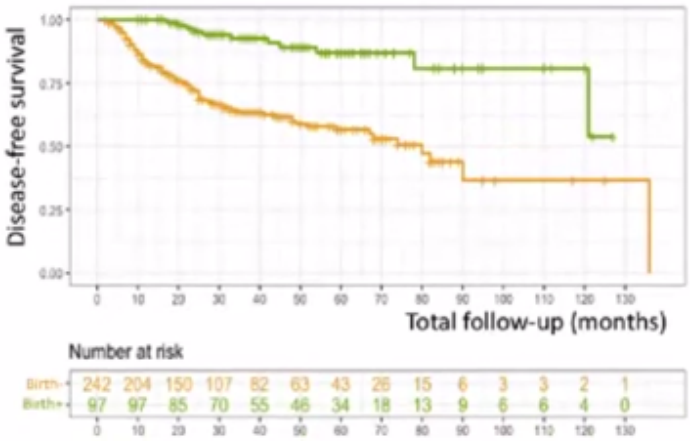
	n	Events (%)	HR (95% CI)	Log Rank P value
ART+	89	23 (25.8%)	0.80 (0.50-1.27)	0.3
ART-	250	78 (31.2%)		



Уровень значимости больше 0.03, что говорит об одинаковости групп. Это соответствует мнению врачей: операция EKO не влияет на рецидив

C.

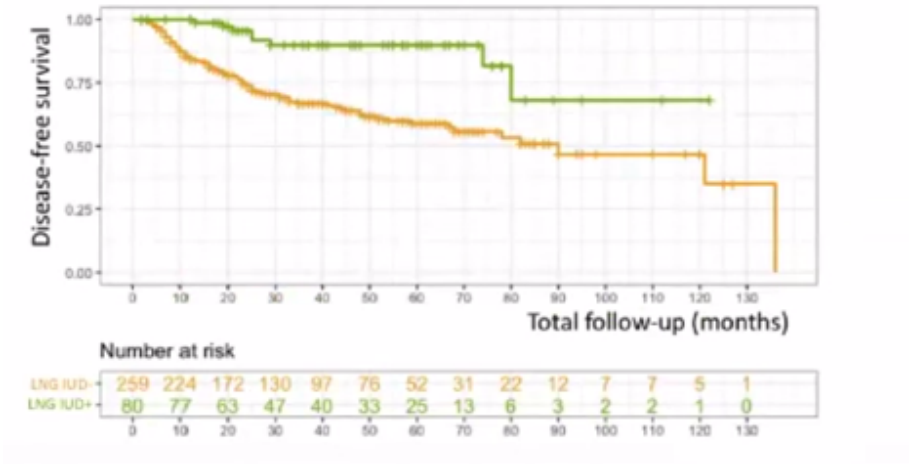
	n	Events (%)	HR (95% CI)	Log Rank P value
Birth+	97	11 (11.3%)	0.21 (0.11-0.40)	0.0001
Birth-	242	90 (37.2%)		



Очень хороший уровень значимости. Лучше родить ребенка

D.

	n	Events (%)	HR (95% CI)	Log Rank P value
LNG IUD+	80	8 (10%)	0.25 (0.12-0.51)	0.0001
LNG IUD-	259	93 (35.9%)		



Number at risk

LNG IUD-	259	224	172	130	97	76	52	31	22	12	7	7	5	1
LNG IUD+	80	77	63	47	40	33	25	13	6	3	2	2	1	0
	0	10	20	30	40	50	60	70	80	90	100	110	120	130

Использование дополнительного лечения уменьшает вероятность рецидива. Это соответствует мнению врачей.

На этом курс закончен!